JUSTIN
Jurnal Sistem dan Teknologi Informasi
p-ISSN : 2460-3562 / e-ISSN : 2620-8989
DOI: 10.26418/justin.v10i1.44103
Vol. 10, No. 1, Januari 2021

Model Prediksi Penderita HCC Menggunakan Algoritma Random Forest

Cindy Wungkana[#1], Megalita Aror[#2], Green Arther Sandag[#3]
[#]*Jurusan Informatika, Fakultas Ilmu Komputer, Universitas Klabat*
Jl. Arnold Mononutu, Airmadidi, Airmadidi Bawah, Kecamatan Airmadidi
[1]s21710041@student.unklab.ac.id
[2]s21630007@student.unklab.ac.id
[3]greensandag@unklab.ac.id

Abstrak

Hepatocellular carcinoma (HCC) atau kanker hati adalah salah satu dari kanker yang paling umum dan menjadi penyebab utama kematian di negara-negara Asia. Presentasi HCC telah berkembang secara signifikan selama beberapa dekade terakhir. Rokok dan minuman beralkohol yang kita konsumsi diketahui menjadi faktor yang mempengaruhi tingkat kehidupan pasien HCC. Penelitian bertujuan untuk mengkaji klasifikasi tingkat kehidupan pasien HCC dengan menggunakan algoritma Random Forest. Dasar dari kriteria penunjang adalah dengan membandingkan algoritma Random Forest dengan algoritma yang lain seperti K-Nearest Neighbors dan Logistic Regression. Percobaan disusun secara teratur dengan mengukur accuracy, precision, recall, dengan rumus yang berhasil dibuat oleh peneliti melalui Google Colaboratory. Hasil percobaan menyatakan bahwa algoritma Random Forest cocok digunakan dalam penelitian ini dengan memiliki accuracy sebesar 100% , recall dan precision sebesar 100% karena berhasil menampilkan performa terbaik.
Kata kunci: HCC, Random Forest, survival rates

Prediction Model for HCC Patients Using the Random Forest Algorithm

Abstract

Hepatocellular carcinoma (HCC) or liver cancer is one of the most common cancers and a leading cause of death in Asian countries. The presentation of HCC has evolved significantly over the last few decades. The cigarettes and alcoholic drinks we consume are known to be factors that affect the life expectancy of HCC patients. This study aims to examine the classification of the life level of HCC patients using the Random Forest algorithm. The basis of the supporting criteria is to compare the Random Forest algorithm with other algorithms such as K-Nearest Neighbors and Logistic Regression. Experiments are arranged regularly by measuring accuracy, precision, recall, with formulas that have been successfully created by researchers through Google Collaboratory. The experimental results show that the Random Forest algorithm is suitable for use in this study with 100% accuracy, 100% recall and precision because it successfully displays the best performance. The results of the experiment show that the Random Forest algorithm is suitable for use in this study because it manages to show the best performance.

Keywords: HCC, Random Forest, survival rates

I. PENDAHULUAN

Dalam kehidupan manusia, hal kesehatan sangat diharapkan dan dijadikan sesuatu yang penting dan diinginkan oleh semua manusia, dibalik kesehatan yang ada tentunya ada beberapa penyakit yang timbul dan diderita oleh beberapa orang. Dalam hal ini, penyakit yang dimaksudkan berhubungan dengan hati, dimana hati merupakan organ yang bermanfaat bagi manusia untuk mencerna segala makanan yang masuk dan harus dijaga.

HCC (*Hepatocellular Carcinoma*) atau lebih sering dikenal dengan istilah kanker hati dimulai dari sel-sel organ hati yang berada di bawah paru-paru sebelah kanan di bawah tulang rusuk, dimana memiliki peranan yang penting seperti mengeluarkan racun dari tubuh. Dibalik fungsi dari hati yang dapat memberikan manfaat bagi manusia, ternyata dapat juga terjadi tumbuhnya sel-sel yang tidak terkendali yang kemudian akan muncul sel tumor yang mengganggu fungsi sel-sel sehat yang ada di hati[1].

Di seluruh dunia, *HCC* (*Hepatocellular Carcinoma*) adalah penyakit yang paling umum keenam dan urutan kedua penyebab kematian karena kanker hati sehingga kasus penyakit ini sekitar 85% banyak terjadi di negara berkembang. Insiden *HCC* (*Hepatocellular Carcinoma*) di Amerika Serikat berubah secara radikal dalam 40 tahun terakhir: tahun 1973 insiden *HCC* adalah 1,51 kasus per

Submitted 11-01-2021; Revised 19-04-2021; Accepted 07-01-2022

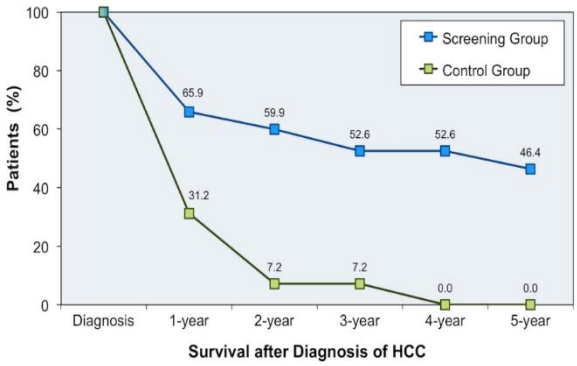
100.000 orang, kemudian meningkat menjadi 6,20 kasus per 100.000 orang pada 2011. Di tahun 2012, diperkirakan ada 24.696 kasus baru *HCC* (*Hepatocellular Carcinoma*) yang didiagnosis [2].

Kanker hati ini adalah kanker kelima yang paling umum pada pria dan ketujuh pada wanita[2]. Berdasarkan *survey* yang dilakukan di Hongkong, tercatat sekitar 1.800 kasus baru yang terjadi setiap tahunnya. Dari kasus baru tersebut, 75% di antaranya adalah pria dengan rata-rata kemunculan penyakit pada usia 63 hingga 69 tahun[3]. Sedangkan hasil survey dari World Health Organization pada tahun 2018 mencatat terdapat 782.000 pasien penderita *HCC* (*Hepatocellular Carcinoma*) yang meninggal[4].

Penelitian oleh *Nadhim, Suharti* dan *Hardian* mengatakan bahwa jenis kelamin laki-laki lebih banyak mengalami penyakit ini daripada perempuan dengan rasionya 4,4 : 1 yang dilakukan penelitiannya pada RSUP. Dr. Kariadi di Semarang.

Menurut dr. Rusdina Bte Ladju, Ph.D mengatakan bahwa penyakit liver terjadi karena adanya *Hepatocellular Carcinoma* yang 99% adanya sirosis dan radang pada hati sehingga diprediksi pada tahun 2040, penyakit kanker ini akan terus meningkat dan faktor terjadinya penyakit ini karena gaya hidup yang tidak sehat[5].

Orang yang tidak memiliki gejala *HCC* (*Hepatocellular Carcinoma*) harus menjalani skrining pasien dimana akan dilakukan pengujian yang walaupun tidak menunjukkan tanda-tanda terkena penyakit ini namun mereka harus dilakukan pengawasan. Gambar dibawah ini akan menunjukkan dampak dari proses skrining yang dilakukan yang dapat membuat keberlangsungan hidup tetap bertahan.



Gambar 1. Dampak Skrining terhadap kelangsungan hidup setelah diagnosis HCC

Pada gambar 1 dijelaskan bahwa pasien dengan hepatitis virus kronis yang menjalani skrining untuk HCC telah meningkatkan kelangsungan hidup setelah diagnosis *HCC* jika dibandingkan dengan kelompok kontrol yang tidak menerima skrining untuk *HCC* [2]. Dapat disimpulkan bahwa orang dengan kelompok control tidak dapat bertahan hidup karena mereka tidak melakukan proses skrining sehingga tidak ada pengawasan dan penyakit kanker mereka akan diketahui jika sudah kronis.

Tingkat bertahan hidup seseorang menurut *National Cancer Institute Surveillance, Epidemiology and End Results* (SIER) dengan mengambil data pasien yang terkena kanker hati, dimana lebih awal didiagnosis maka akan lebih baik untuk mencegah seseorang akan bertahan hidup dari penyakit yang diderita daripada sudah didiagnosa ketika sudah mendapat kanker hati dan sudah ada di tahap akhir.

Terdapat beberapa faktor yang mempengaruhi tingkat kehidupan pasien *HCC* (*Hepatocellular Carcinoma*) seperti jumlah berapa banyak rokok dan minuman beralkohol yang dikonsumsi[6]. Namun dari faktor-faktor tersebut masih belum diketahui mana yang paling dominan terhadap kematian pasien penderita *HCC* (*Hepatocellular Carcinoma*). Penyakit ini relatif lebih sulit untuk disembuhkan karena pasien penderita kanker hati biasanya terdiagnosis pada stadium menengah atau akhir[7].

Penelitian ini dilakukan untuk memprediksi dan mengetahui tingkat kehidupan pasien apakah bertahan atau tidak setelah didiagnosa menderita penyakit berdasarkan 2 faktor utama; jumlah bungkus rokok yang dikonsumsi setiap tahuhn dan banyaknya gram dari *alcohol* yang dikonsumsi perhari, yang telah diuji menggunakan algoritma *Random Forest*. Algoritma ini digunakan karena jumlah data dari *dataset* penelitian yang dilakukan memiliki jumlah yang besar. Pun sangat cocok karena menghasilkan jumlah error yang rendah dan dapat mengatasi jumlah training yang besar karena data yang digunakan bukan biner tetapi non-biner [8].

Penelitian yang dilakukan oleh *Tuasikal* dan *Widodo* untuk prediksi terhadap penyakit kanker payudara menggunakan algoritma *Random Forest*, memiliki tingkat akurasi sebesar 100% dibandingkan menggunakan algoritma yang dipakai pada penelitian mereka yaitu *Support Vector Machine (SVM)* yang hanya 94%. Dapat dikatakan bahwa algoritma Random Forest lebih baik untuk akurasi dan digunakan pada data yang besar yang ada pada penelitian *Tuasikal* dan *Widodo* dengan mengambil data dari pasien yang terkena kanker payudara di *UCI Machine Learning Wicoxsin University.*

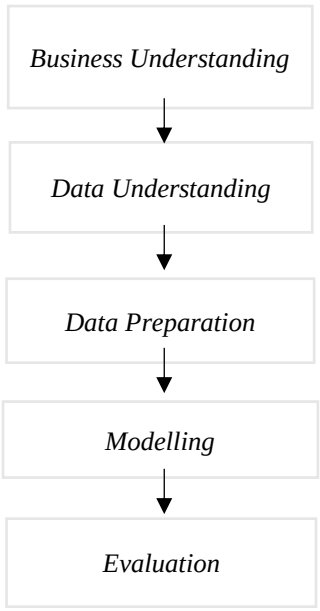
Diharapkan penelitian ini dapat bermanfaat untuk memberikan informasi kepada pembaca khususnya tenaga medis untuk memberikan penanganan lebih khusus kepada penderita *HCC.*

II. METODE PENELITIAN

Pada bagian ini akan dijelaskan bagaimana metode dan apa saja yang dilakukan untuk mencapai tujuan dari penelitian yang akan dilakukan.

A. Desain Penelitian

Strategi yang dipilih untuk mengintegrasikan penelitian ini adalah dengan menggunakan fase *CRISP – DM* (*Cross-Industry Standard Process for Data Mining).* Pada Gambar 2 terdapat fase CRISP-DM yang digunakan untuk menganalisis penelitian ini dengan penjelasan dari tiap-tiap bagian dalam fase CRISP-DM ini.



Gambar 2. Fase *CRISP-DM*

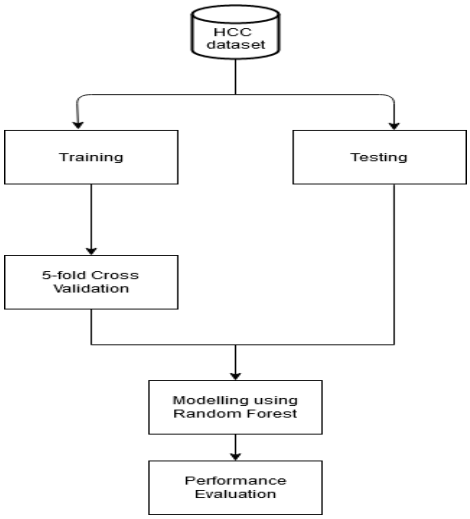
B. Fase Pemahaman Bisnis (Business Understanding Phase)

Pada tahap awal ini, peneliti menggunakan data dari Kaggle [15]. *Dataset* yang digunakan merupakan dataset terbaru yang di-update pada tahun 2018 Variabel target adalah kelangsungan hidup pada 1 tahun, dan dikodekan sebagai variabel biner: 0 (mati) dan 1 (bertahan). Tingkat ketidakseimbangan kelas tertentu juga ada (63 kasus diberi label sebagai "meninggal" dan 102 sebagai "kehidupan"). Dari dataset tersebut peneliti memutuskan untuk memprediksi tingkat keberlangsungan hidup penderita penyakit HCC dengan menggunakan algoritma Random Forest dan akan dibandingkan dengen beberapa algoritma yang lain.

C. Fase Pemahaman Data (Data Understanding Phase)

Pada tahap kedua ini, peneliti mencari pemahaman tentang dataset. Pengumpulan data kami menggunakan *Hepatocellular Carcinoma dataset* yang ada di *Kaggle* [15]. *Dataset* yang ada diperoleh dari Rumah Sakit Universitas yang ada di Portugal dan berisi data pasien sebanyak 165 yang didiagnosis menderita penyakit *HCC* (*Hepatocellular Carcinoma)* secara nyata. Teknologi terkini dalam pengelolaan *HCC* menurut Pedoman Praktik Klinis *EASL-EORTC* (Asosiasi Eropa untuk Studi Hati - Organisasi Eropa untuk Penelitian dan Pengobatan Kanker) memiliki *dataset* berisi 49 fitur yang dipilih. Pada penelitian ini, kami menggunakan *Hepatocellular Carcinoma dataset* yang ada di *Kaggle* [13]. *Dataset* yang digunakan ini memiliki 204 *rows* dan 50 *column*, dimana ini terdiri dari 49 *attributes* independent dan 1 *attribute dependent* yaitu class.

D. Fase Pengolahan Data (Data Preparation Phase)



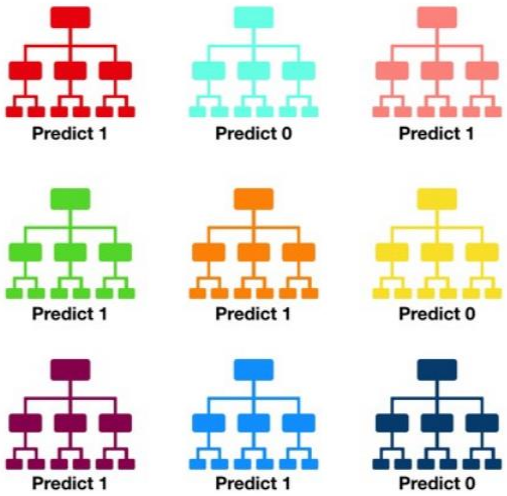
Gambar 3. Arsitektur Untuk Klasifikasi Tingkat Kehidupan Pasien *HCC*

Sebelum masuk pada tahap *modelling* yang akan dilakukan dengan menggunakan *Explaratory Data Analysis* (*EDA*), pada tahap *data preparation* ini peneliti melakukan pencarian *missing values* dan pembersihan data, dimana data yang tidak dapat digunakan akan dibuang. Peneliti melakukan persiapan data mencakup semua kegiatan untuk menyusun *dataset* akhir dari data mentah awal guna menyiapkan data untuk diproses lebih lanjut. Gambar 3 menjelaskan proses klasifikasi tingkat kehidupan pasien *HCC* yang dimulai dengan pengambilan *Hepatocellular Carcinoma Dataset* dari website *Kaggle*, dengan *dataset* yang terdiri dari 204 data ke dalam dua bagian yaitu training sebesar 80% yang berjumlah 163 dan testing sebesar 20% yang berjumlah 40.

E. Fase Permodelan (Modelling Phase)

Pada tahap ke empat, peneliti membuat model berdasarkan beberapa teknik pemodelan yang berbeda. Di sini peneliti telah memilih beberapa algoritma, di antaranya; *Random Forest*, *K-Nearest Neighbors*, *Naive Bayes*, dan *Logistic Regression*. Setelah itu peneliti membagi data menjadi set pelatihan, pengujian, dan validasi. Data yang didapatkan langsung dibandingkan dengan algoritma lain supaya peneliti dapat menafsirkan hasil model berdasarkan pengetahuan umum, kriteria keberhasilan yang telah ditentukan sebelumnya, dan desain pengujian.

F. Algoritma Random Forest



Gambar 4. Visualisasi *Random Forest* [16]

Random Forest mampu mengklasifiksi data yang memiliki atribut tidak lengkap, dapat digunakan untuk klasifikasi dan regresi, sehingga menghasilkan akurasi yang lebih tinggi dan dapat mengatasi jumlah data besar secara efisien. Metode ini digunakan untuk menciptakan *Desicion Tree* yang terdiri dari *root node, internal node, dan leaf node* dengan mengambil atribut dan data secara acak mengikuti ketentuan yang diberlakukan.
Desicion Tree dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain*. Untuk menghitung nilai *entropy* digunakan rumus sebagai berikut [18]:

$$Entropy\ (Y)\ =\ -\sum p\ (c|Y)\ i\ log2\ p\ (c|Y)\text{.....(1)}$$

Keterangan:
Y = himpunan kasus.
p(c|Y) = proporsi nilai Y terhadap kelas c.

$$Gain\ (T,X)\ =\ Entropy(T)\ -\ Entropy(T,X)\text{,...(2)}$$

Keterangan:
T = Variabel Target
X = Fitur yang akan dibagi
Entrophy (T,X) = Entropi dihitung setelah data dipisahkan pada fitur X.

G. Fase Evaluasi (Evaluation Phase)

Untuk fase ini setelah kami melakukan pengujian dengan melakukan *training* dan *testing* terhadap data yang ada maka hasilnya dari 49 atribut yang ada, beberapa dari itu memiliki pengaruh terhadap apa yang membuat seseorang penderita *Hepatocellular Carcinoma* dapat bertahan atau tidak dengan melihat faktor-faktor yang penting dari hasil yang telah dimodelkan. Fase evaluasi akan menguji faktor-faktor yang ada dengan menggunakan *Feature Importance* dan penggunaan algoritma Random Forest beserta algoritma pembanding lainnya dengan menggunakan *Performance Evaluation* dan *5-fold Cross Validation*.

a) *Feature Importance*

Digunakan mencari atribut mana yang paling berpengaruh terhadap model prediksi. Jika atribut tersebut mengalami perubahan, maka akan berpengaruh pada model prediksi [17]. Feature importance menggunakan pengujian dengan signifikasi statistik untuk memutuskan mana atribut yang mempengaruhi atribut yang akan diprediksi [18]. Atribut yang dimaksud yaitu penggunaan rokok setiap tahun dan penggunaan alcohol setiap hari.

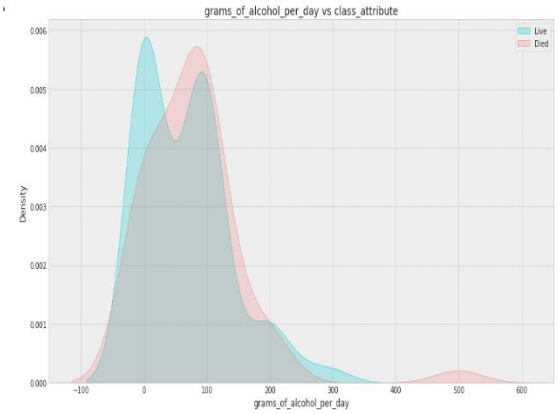
b) *Performance Evaluation*

Untuk pengukuran performance, ada beberapa hal yang diukur berupa *Accuracy*, *Precission*, *Recall*, dan *Specificity* [18].

Accuracy, mengukur rasio prediksi yang benar, baik itu positif dan negatif pada keseluruhan data.

$$Accuracy\ =\ (TP\ +\ TN\ )\ /\ (TP + FP + FN + TN)\text{..(3)}$$

Precission, mengukur rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif.



Gambar 4. Histogram penggunaan alkohol per gram setiap hari

$$Precission\ =(TP)\ /\ (TP + FP)\text{,.....(4)}$$

Recall (Sensitifitas), rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif.

$$Recall\ =\ (TP)\ /\ (TP\ +\ FN)\text{,.....(5)}$$

RMSE (Root Mean Square Error, merupakan salah satu cara untuk mengevaluasi model regresi linear dengan mengukur tingkat akurasi hasil perkiraan suatu model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}\text{,.....(6)}$$

c) *5-Fold Cross Validation*

Cross Validation adalah sebuah teknik yang digunakan dalam pengevaluasian model dengan cara membagi sampel dari *dataset* asli ke dalam *training set* untuk melatih model dan *testing set* untuk melakukan evaluasi terhadap model tersebut. *Dataset* yang telah melalui proses *feature extraction* kemudian dibagi ke dalam dua proses evaluasi,

yang kemudian akan diprediksi menggunakan algoritma *Random Forest.*

H. Fase Penyebaran (Deployment Phase)

Pada bagian ini akan didapat pengetahuan dari hasil evaluasi yang ada dimana akan menjadi sebuah penjelasan dari tujuan yang diharapkan dimana *accuracy* dari algoritma *Random Forest* yang dipilih memiliki kinerja yang cukup baik dan faktor-faktor yang ada sebelumnya akan diseleksi dan didapati faktor penting yang mempengaruhi penelitian yang dilakukan ini.

III. HASIL DAN PEMBAHASAN

Pada bab ini, akan menjelaskan hasil dari pengujian algoritma *Random Forest* terhadap penelitian yang dilakukan dengan menggunakan *Explaratory Data Analysis* yang dapat menganalisis data untuk menghapus kumpulan data dan kemudian menerapkan teknik pembelajaran mesin. Kami menggunakan ini untuk mengimplementasikan uji coba algoritma *Random Forest* yang akan dibuat untuk menghasilkan sebuah model dan faktor-faktor yang berpengaruh dalam tingkat kehidupan pasien *HCC*.

A. Explaratory Data Analysis

a) Penggunaan Alkohol Setiap Hari pada Pasien

Pada bagian ini peneliti ingin menampilkan bagaimana penggunaan alkohol setiap hari pada pasien penderita *HCC* yang bertahan hidup. Hasil dari Gambar 5 ditujukkan dengan plot yang memperlihatkan hasil dari berapa gram untuk konsumsi alkohol agar sehingga dapat bertahan hidup. Dan dari hasil yang ada dapat disimpulkan bahwa sekitar <100-gram alkohol dapat bertahan hidup sedangkan saat 100-gram alkohol dikonsumsi banyak yang tidak dapat bertahan hidup.

b) Penggunaan Rokok Setiap Tahun pada Pasien

Pada bagian ini peneliti ingin menjelaskan bagaimana penggunaan rokok per bungkus setiap tahun pada pasien penderita *HCC* yang bertahan hidup. Hasil pada Gambar 5 menunjukkan bahwa orang yang bertahan hidup mengkonsumsi sekitar 20 bungkus rokok per tahun sedangkan orang yang >20 bungkus rokok tidak dapat bertahan hidup.

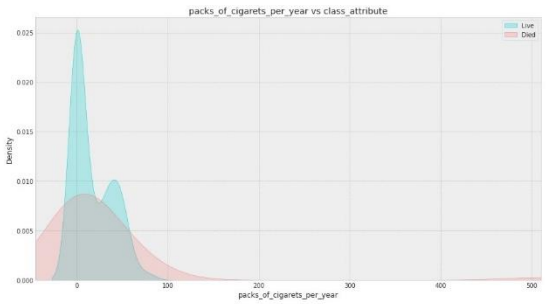
B. Perbandingan Algoritma dengan Independent Dataset

Setelah hasil yang ditampilkan pada Tabel 1 disimpulkan bahwa *accuracy* dari prediksi menggunakan algoritma *Random Forest* adalah 100%, begitu juga dengan *precision* dan *recall* yang memiliki nilai sama yaitu 100%. Adapun beberapa algoritma-algoritma yang digunakan sebagai pembanding untuk melihat mana algoritma yang baik, dilihat dari tingkat akurasinya dan dimodelkan dalam bentuk Tabel 1 untuk perhitungan dan perbandingan Algoritma *Random Forest* dengan Algoritma *Logistic Regression*, *K-Nearest Neighbors*, dan *Naïve Bayes* dengan menggunakan *Independent Test.*

C. Perbandingan Algoritma dengan *Cross Valdiation*

TABEL II
HASIL KOMPARASI ALGORITMA *RANDOM FOREST* DENGAN ALGORITMA LAINNYA MENGGUNAKAN *5-FOLD VALIDATION*

Hasil *Cross Validation*			
Algorithm	***Accuracy (%)***	***Recall (%)***	***Precision (%)***
Random Forest	100%	100%	100%
K-Nearest Neighbors	84.29%	82.00%	83.00%
Logistic Regression	97.58%	97.00%	97.00%
Naïve Bayes	99.53%	82.00%	83.00%



Gambar 5. Histogram Penggunaan Rokok Setiap Tahun

Tabel 2 menunjukkan hasil ketika menggunakan *Cross Validation* pada algoritma *Random Forest.* Pada penelitian ini juga kami menggunakan *5-Fold Cross Validation* yang lebih menunjukkan keseimbangan antara *precision* dan *recall* dan lebih cocok digunakan dalam penelitian kami karena dapat menaikkan jumlah *precision* dan *recall* lebih tinggi lagi sehingga mendapat model terbaik dibandingkan hanya mengandalakan akurasi saja. Dapat dilihat juga bahwa hasil menggunakan *Cross Validation* adalah 100% bersamaan dengan nilai *accuracy*, *precision* dan *recall* yang juga memiliki nilai akhir yang sama.

Adapun beberapa algoritma-algoritma yang digunakan sebagai pembanding untuk melihat mana algoritma yang baik, dilihat dari tingkat akurasinya dan dimodelkan dalam bentuk Tabel 2 untuk perhitungan dan perbandingan Algoritma *Random Forest* dengan Algoritma *Logistic Regression*, *K-Nearest Neighbors*, dan *Naïve Bayes* dengan menggunakan *Cross Validation*.

IV. KESIMPULAN

Berdasarkan pembahasan dan analisis dari penelitian ini yang sudah kami tulis, diperoleh kesimpulan bahwa hasil prediksi kami menggunakan metode CRISP-DM dan untuk modelling menggunakan algoritma Random Forest sangat cocok digunakan pada penelitian kami karena modelnya memang sangat bagus sehingga accuracy, precision dan recall mencapai 100% pada Independent Test begitu juga

ketika menggunakan Cross Validation mendapat 100% dan didukung juga oleh penelitian dari Aliady14], yang memiliki accuracy yang baik sebesar 94,5%, untuk itu Random Forest ini merupakan algoritma yang tepat dan sangat baik dalam penggunaannya. Penelitian ini juga meneliti 2 parameter yang mempengaruhi tingkat kehidupan para penderita HCC untuk bertahan hidup atau tidak, dengan melihat dari faktor penggunaan rokok dan faktor penggunaan alkohol.

SARAN

Berdasarkan pembahasan dan analisis dari penelitian ini yang sudah kami tulis, diperoleh saran yang dapat berguna bagi peneliti selanjutnya seperti berikut: 1. Diharapkan untuk penelitian selanjutnya diberikan saran bahwa penggunaan Algoritma Random Forest sangat baik untuk digunakan pada dataset yang besar karena memiliki tingkat akurasi yang baik dibandingkan algoritma lainnya. 2. Diharapkan pada peneliti selanjutnya untuk mengimplementasi Algoritma Random Forest untuk use case yang berbeda

DAFTAR PUSTAKA

[1] "Gejala Kanker Hati Stadium Awal (Harus Diwaspadai) - Dokter Sehat", *Doktersehat.com*,2019. [Online]. Available: https://doktersehat.com/gejala-kanker-hati-stadium-awal/. [Accessed: 18- Apr-2020].

[2] [2] Rena. K. Fox, "*Surveillance for Hepatocellular Carcinoma*" - ", *Hepatitis C Online*,31-May-2018. [Online]. Available: https://www.hepatitisc.uw.edu/go/evaluation-staging-monitoring/surveillance-hepatocellular-carcinoma/core-concept/all#page-title [Accessed: 21- Apr- 2020].

[3] [3] *Www21.ha.org.hk*,2019.[Online]. Available: https://www21.ha.org.hk/smartpatient/EM/MediaLibraries/EM/Diseases/Cancer/Liver%20Cancer/Cancer-Liver-Cancer-Indonesian.pdf?ext=.pdf. [Accessed: 21-Apr- 2020].

[4] [4] "New Global Cancer Data: GLOBOCAN 2018 | UICC", *Uicc.org*, 2019. [Online]. Available: https://www.uicc.org/news/new-global-cancer-data-globocan-2018. [Accessed: 21- Apr- 2020].

[5] [5] Makassarmetro.com, "Dosen FK Unhas Paparkan Hasil Penelitian Terbaru tentang Kanker Hati dan Malaria," *Makassarmetro.com*, 15-Nov-2019. [Online]. Available: https://makassarmetro.com/2019/11/15/dosen-fk-unhas-paparkan-hasil-penelitian-terbaru-tentang-kanker-hati-dan-malaria. [Accessed: 24-Apr-2020].

[6] [6] Bosetti C, "Hepatocellular carcinoma epidemiology. - PubMed - NCBI", *Ncbi.nlm.nih.gov*, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25260306. [Accessed: 18-Apr-2020].

[7] [7] K. Media, "Waspadai Kanker Hati", *KOMPAS.com*, 2019. [Online]. Available: https://nasional.kompas.com/read/2011/02/01/18410953/waspadai.kanker.hati. [Accessed: 18-Apr-2020].

[8] [8] N. Azizah, IMPLEMENTASI DAN ANALISA WAKTU KOMPUTASI PADA ALGORITMA RANDOM FOREST DENGAN PARALLEL COMPUTING DI R. Universitas Pendidikan Indonesia, 2017.

[9] [9] K. Media, "Apa Saja Tanda Kanker Hati?", *KOMPAS.com*, 2019. [Online]. Available: https://lifestyle.kompas.com/read/2014/08/27/154606723/Apa.Saja.Tanda.Kanker.Hati. [Accessed: 21-Apr-2020].

[10] [10] "Sulit Dideteksi Dini, Kenali Ragam Gejala Kanker Hati", *suara.com*, 2019. [Online]. Available: https://www.suara.com/health/2019/03/26/141248/sulit-dideteksi-dini-kenali-ragam-gejala-kanker-hati?page=all. [Accessed: 18-Apr-2020].

[11] [11] M. Khalilia, S. Chakraborty, and M. Popescu, "Predicting disease risks from highly imbalanced data using random forest," *BMC Medical Informatics and Decision Making*, vol. 11, no. 1, 2011.

[12] [12] "Prediction of hepatocellular carcinoma patient survival using machine learning classification rules. | Journal of Clinical Oncology", *Ascopubs.org*, 2019. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/JCO.2019.37.15_suppl.e15649. [Accessed: 18-Apr-2020].

[13] [13] G.A. Sandag, N.E. Tedry, S. Lolong "Classification of Lower Back Pain Using K-Nearest Neighbor Algorithm". Cyber and IT Service Management (CITSM), Agustus 2018.

[14] [14] H. Aliady, N. J. Tuasikal, and E. Widodo, "IMPLEMENTASI SUPPORT VECTOR MACHINE (SVM) DAN RANDOM FOREST PADA DIAGNOSIS KANKER PAYUDARA." Seminar Nasional Teknologi Informasi dan Komunikasi 2018 (SENTIKA 2018), Yogyakarta, 23-Mar-2018.

[15] [15] "HCC dataset", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/mrsantos/hcc-dataset. [Accessed: 20- Nov- 2019].

[16] [16] "Understanding Random Forest", *Medium*, 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed: 27- Nov- 2019].

[17] [17] C. Molnar, "5.5 Permutation Feature Importance | Interpretable Machine Learning", *Christophm.github.io*, 2019. [Online]. Available: https://christophm.github.io/interpretable-ml-book/feature-importance.html. [Accessed: 28- Nov- 2019].

[18] [18] "Feature importance", *Bayesserver.com*, 2019. [Online]. Available: https://www.bayesserver.com/docs/learning/feature-selection. [Accessed: 28- Nov- 2019].